DETAILED ACTION

Status of Claims

1. This action is responsive to Appeal Brief filed on January 9, 2009 where claims 1-18 were pending.

EXAMINER'S AMENDMENT

2. An examiner's amendment to the record appears below. Should the changes and/or additions be unacceptable to applicant, an amendment may be filed as provided by 37 CFR 1.312. To ensure consideration of such an amendment, it MUST be submitted no later than the payment of the issue fee.

Authorization for this examiner's amendment was given in a telephone interview with Philip Lyren (reg 40709) on 4/9/09.

The application has been amended as follows:

SEE ATTACHED LISTING OF CLAIMS

Allowable Subject Matter

3. Claims 1-6,8,9,11-13,15-17 are allowed.

4. The following is an examiner's statement of reasons for allowance: Applicants invention of allocating spare server resources, is found to be patentable. Prior art references found to be pertinent to Applicants disclosure (such as Patent Numbers: 7451183, 6065062, 6226377, and US Patent Publication Nos. 20050033794, 20040205219, 20040010544), either only teach minor aspects of the invention or only teach the general

environment of the invention. The collective prior art, either singly or in combination, do not teach the claim limitations.

The primary reason for allowance is the novelty of using a non-iterative queuing model for predicting the average response time for a server system in response to measured arrival rates of transaction requests into each of two scaleable tiers of server machines, an average service demand at each of the two scaleable tiers of server machines, and a number of servers allocated to each of the two scaleable tiers of server machines. And based on this average response time increasing a number of server machines processing transactions for each of the two scaleable tiers of server machines by allocating the spare server machines to process a portion of the transactions.

Any comments considered necessary by applicant must be submitted no later than the payment of the issue fee and, to avoid processing delays, should preferably accompany the issue fee. Such submissions should be clearly labeled “Comments on Statement of Reasons for Allowance.”

Conclusion

5. The prior art made of record and not relied upon is considered pertinent to applicant's disclosure. Particularly:

US Publication 2005/0033794, where Aridor et al discloses managing multitier application complexes. Aridor teaches detecting excess load on a second tier of servers, determining available servers from a server pool, and requesting additional servers from the server pool (see paragraph 16).

US Patent 7451183, Romero et al discloses balancing processors in a partitioned server. Romero teaches a partition requiring an extra processor, deactivating an active processor from another partition, and then activating a reserve processor for the partition requiring an extra processor (see Abstract).

Any inquiry concerning this communication or earlier communications from the examiner should be directed to RAMY M. OSMAN whose telephone number is (571)272-4008. The examiner can normally be reached on M-F 9-5.

If attempts to reach the examiner by telephone are unsuccessful, the examiner's supervisor, Ario Etienne can be reached on (571) 272-4001. The fax phone number for the organization where this application or proceeding is assigned is 571-273-8300.

Information regarding the status of an application may be obtained from the Patent Application Information Retrieval (PAIR) system. Status information for published applications may be obtained from either Private PAIR or Public PAIR. Status information for unpublished applications is available through Private PAIR only. For more information about the PAIR system, see http://pair-direct.uspto.gov. Should you have questions on access to the Private PAIR system, contact the Electronic Business Center (EBC) at 866-217-9197 (toll-free). If you would like assistance from a USPTO Customer Service Representative or access to the automated information system, call 800-786-9199 (IN USA OR CANADA) or 571-272-1000.

/Ramy M Osman/ April 12, 2009
Primary Examiner, Art Unit 2457

LISTING OF CLAIMS

1. (currently amended) A server system comprising:

at least two scaleable tiers of server machines;

a server pool including plural spare server machines;

means for computing an average response time for the server system to respond to transaction requests at the two scaleable tiers of server machines; and

means for increasing a number of server machines processing transactions for each of the two scaleable tiers of server machines by allocating the spare server machines to process a portion of the transactions, wherein the spare server machines are allocated to process a portion of the transactions when the average response time for the server system to respond to the transaction requests is greater than or equal to a specified average response time, wherein said means for computing further comprises a non-iterative queuing model for predicting the average response time for the server system in response to measured arrival rates of transaction requests into each of the two scaleable tiers of server machines, an average service demand at each of the two scaleable tiers of server machines, and a number of servers allocated to each of the two scaleable tiers of server machines.

2. (previously presented) The server system of claim 1 further comprising means for determining costs associated with allocating the number of server machines at each of the two scaleable tiers of server machines.

3. (previously presented) The server system of claim 2 wherein said means for determining further comprises means for minimizing costs associated with allocating an optimized number of server machines at each of the two scaleable tiers of server machines.

4. (previously presented) The server system of claim 3 wherein said means for minimizing comprises:

means operatively coupled to said server system for receiving input parameters and for solving:

$$\sqrt{\gamma} = \frac{\sum_{i=1}^{n} \sqrt{h_i s_i u_i}}{T - \sum_{i=1}^{n} s_i};$$

where: γ is a shadow price of the average response time; $h_1, h_2, \ldots h_n$ are weights reflecting a cost of different types of servers located at each of the two scaleable tiers of server machines; s is an average service time; u is a measured average utilization rate expressed in a single-machine percentage; and T is the average response time.

5. (previously presented) The server system of claim 1, wherein the average response time is determined by examining a time that the transaction requests are pending at each of the two scaleable tiers of server machines.

6. (previously presented) The server system of claim 1 further comprising:

a contractual relationship between a system operator and at least one contracting party; and

means for adjusting prices charged by said system operator to at least one third party in response to a change in an allocation of server machines in each of the two scaleable tiers of server machines.

7. (canceled).

8. (currently amended) A method for allocating a server machine to at least two tiers of a server system, said method comprising:

computing, by a computer, an expected average response time as a function of transaction requests and an amount of resources allocated to each of the two tiers of the server system;

determining whether an optimization problem is feasible;

computing a lower bound and an upper bound on a number of server machines at each of the two tiers of said server system required to meet the average response time;

computing a solution specifying a number of server machines allocated to each of the two tiers of said server system;

computing an average time that transaction requests are pending at each of the two tiers; ~~and~~

automatically increasing the number of server machines, from a pool of server machines, allocated to one of the two tiers at a point in time when the average time the transaction requests are pending at the one of the two tiers is greater than or equal to a pre-determined limit; and predicting, with a non-iterative queuing model, an average server system response time in response to measured arrival rates of transaction requests into said two tiers of server machines, an average service demand at said two tiers of server machines; and a number of servers allocated to said two tiers of server machines.

9. (previously presented) The method of claim 8 wherein said computing an expected average response time further comprises:

obtaining at least one input value for an average arrival rate of transaction requests into each of the two tiers of said server system;

obtaining at least one input value for an average service demand at each of the two tiers of said server system; and

obtaining at least one input value for the number of server machines allocated at each of the two tiers of said server system.

10. (canceled).

11. (currently amended) An assembly for allocating server machines in a server system comprising:

at least two tiers of server machines;

a pool of spare server machines that process transactions for the two tiers of server machines;

means for computing an average response time for said two tiers of server machines to respond to a plurality of transaction requests; and

means for increasing and decreasing a number of server machines from said pool that process transactions for said two tiers of server machines when average response times for processing transactions at the two tiers of server machines exceed a specified average response time, wherein said means for computing further comprises a non-iterative queuing model for predicting an average server system response time in response to measured arrival rates of transaction requests into said two tiers of server machines, an average service demand at said two tiers of server machines, and a number of servers allocated to said two tiers of server machines.

12. (previously presented) The assembly of claim 11, wherein the average response time is determined by examining a time that the transaction requests are pending at the two tiers of server machines.

13. (previously presented) The assembly of claim 11 further comprising:

a contractual relationship between a system operator and at least one contracting party; and

means for adjusting prices charged by said system operator to said at least one contracting party in response to a change in an allocation of server machines in said two tiers of server machines.

14. (canceled).

15. (currently amended) A server system comprising:

an open queuing network of multiple server machines with each server machine having a processor-sharing queue with a single critical resource;

at least two tiers of server machines; and

a computer-readable medium comprising instructions for:

(i) predicting an average system response time of said multiple server machines based on an arrival rate of transaction requests into each of the two tiers of server machines averaged over all transaction request types and a number of server machines allocated at each of the two tiers of server machines;

(ii) solving a mathematical representation of an optimization objective and constraints of said server system;

(iii) determining a number of server machines for each of the two tiers of server machines in response to said predicted the average system response time; ~~and~~

(iv) automatically increasing the number of server machines, from a pool of server machines, processing transactions for each of the two tiers of server machines at a point in time when an average time that transactions requests are pending at the two tiers of server machines exceeds a threshold; and

(v) predicting, with a non-iterative queuing model, an average server system response time in response to measured arrival rates of transaction requests into said two tiers of server machines, an average service demand at said two tiers of server machines, and a number of servers allocated to said two tiers of server machines.

16. (previously presented) The server system of claim 15 wherein said mathematical representation comprises:

a continuous-relaxation model of a mathematical optimization system; and

an iterative bounding procedure.

17. (previously presented) The server system of claim 15 wherein said instructions for determining the number of server machines for each of the two tiers of server machines is in response to a predicted average system response time and at least one service level agreement (SLA) requirement.

18. (canceled).